Domestic tourism and travelling post the national (India) lockdown

Background

"The world has changed" said Warren Buffet aphoristically as he went on to dispose of all the airline stocks owned by Berkshire Hathaway. While moving into a hazy future, it is clear that international travelling will not be the same, at least not in the next few years. Countries may proscribe foreign travel, maybe even reject visas or best case, our attitude towards 'better safe than sorry' will make us circumspect with our travel decisions. But most of us have an inherent wanderlust attitude, to be able to roam, be free, explore and find adventure. It is possible that although our movements will be severely restricted for the foreseeable future, domestic travel will be the new international travel. Of course it is risky, but humans are restive. As a result, with this small hunch in mind, my intention is to cluster as many monumental sites, scenic spots and travel destinations based on several criteria such as visitor football (at least at a time before COVID-19), ratings, popularity but most importantly on the current state of coronavirus in each of its states and/or districts (specifically number of active cases). With that being kept in mind, we could possibly use the findings to suggest the least dicey hotspots for that vacation we all deserve. The target audience is every citizen of India who wishes to travel keeping in mind the current scenario of COVID-19. This could also help travel agencies and local service industries market better.

Data

Approximately 450 touristic destinations were obtained from Google Travel through web scraping. 6 similar URLs were used for different types of touristic destinations like art and culture, amusement parks and beaches. (example of one: https://www.google.com/travel/things-to-do/see-all?g2lb=2502548%2C4215767%2C4258168%2C4260007%2C4270442%2C4274032%2C4291318%2C4305595%2C4306835%2C4317915%2C4328159%2C4329288%2C4333265%2C4358983%2C4366684%2C4367954%2C4369397%2C4372336%2C4373848%2C4380601%2C4270859%2C4284970%2C4291517%2C4316256%2C4356899&hl=en&gl=in&un=1&otf=1&dest_mid=%2Fm%2F03rk0&dest_state_type=sattd&dest_src=ts&tcfs=EgoKCC9tLzAzcmsw&sa=X#ttdm=26.112690_78.025407_5&ttdmf=%252Fm%252F081jv3)

Google Places and Google Geocoding (and reverse Geocoding) APIs were used for obtaining several key data attributes (latitude, longitude, state, city, district, ratings, ratings and number of ratings). Originally intended to use Foursquare, but it is very limited for the Indian reality and specifically for this problem. Wikipedia was used for obtaining district and state wise population and area information (https://en.wikipedia.org/wiki/List_of_districts_in_India) through web scraping. All web scraping was done with the help of the Requests and Beautiful Soup libraries. Lastly, I used the publicly available coronavirus APIs (https://api.covid19india.org/) to get updated information on the number of active cases in both districts as well as states. A publicly available GeoJSON file (https://github.com/Subhash9325/GeoJson-Data-of-Indian-States) for states in India was also obtained for plotting choropleth maps for better visualization.

Methodology

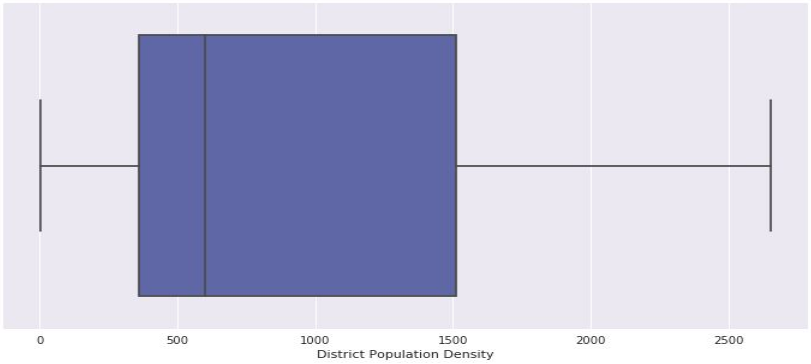
Final dataframe obtained after web scraping, usage of APIs, preparation and cleaning:

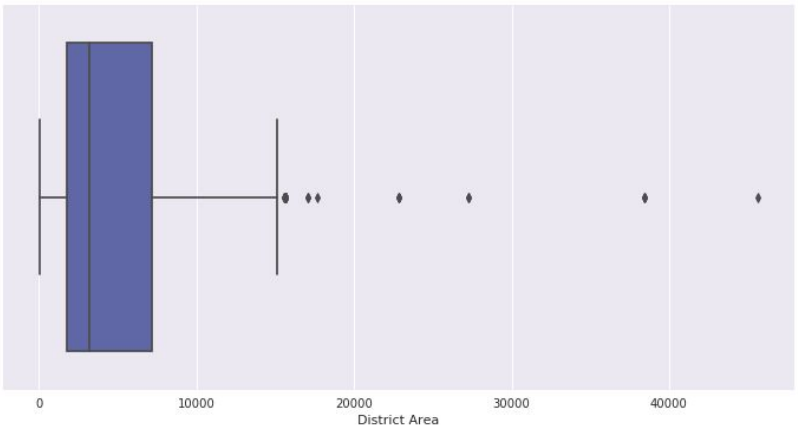
	Name	Latitude	Longitude	District	State	Ratings	Number of Ratings	Type of Destination	District Cases	District Population	District Area
0	Taj Mahal	27.175145	78.042142	Agra	Uttar Pradesh	4.6	167260	[establishment, point_of_interest, tourist_att...	278	4380793	4027.0
1	Amber Palace	26.985487	75.851345	Jaipur	Rajasthan	4.6	93089	[establishment, point_of_interest, premise, to...	586	6663971	11152.0
2	Ranthambore National Park	26.017327	76.502574	Sawai Madhopur	Rajasthan	4.4	3007	[establishment, park, point_of_interest, touri...	8	1338114	4500.0
3	Red Fort	28.656159	77.241020	North Delhi	Delhi	4.4	100257	[establishment, point_of_interest, tourist_att...	60	887978	59.0
4	Hawa Mahal	26.923936	75.826744	Jaipur	Rajasthan	4.4	82997	[establishment, point_of_interest, premise, to...	586	6663971	11152.0
5	City Palace, Jaipur	26.925771	75.823658	Jaipur	Rajasthan	4.4	34553	[establishment, museum, point_of_interest, pre...	586	6663971	11152.0

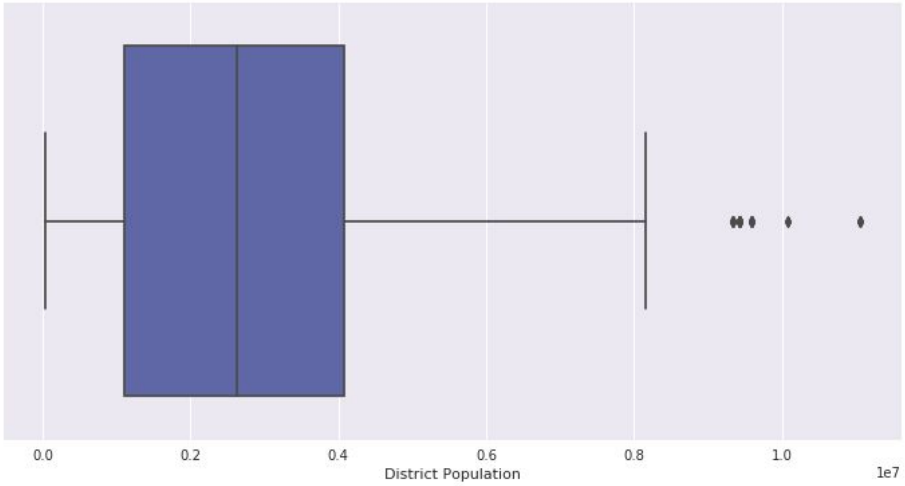
District Cases/Person	District Population Density	District Cases/km2	Clusters_KMeans
0.000063	1087.855227	0.069034	4
0.000088	597.558375	0.052547	4
0.000006	297.358667	0.001778	3
0.000068	15050.474576	1.016949	4
0.000088	597.558375	0.052547	4

I derived some features through basic arithmetic, such as Population density which is Total Population by Total Area. Likewise other features were obtained through the combination of primary features.

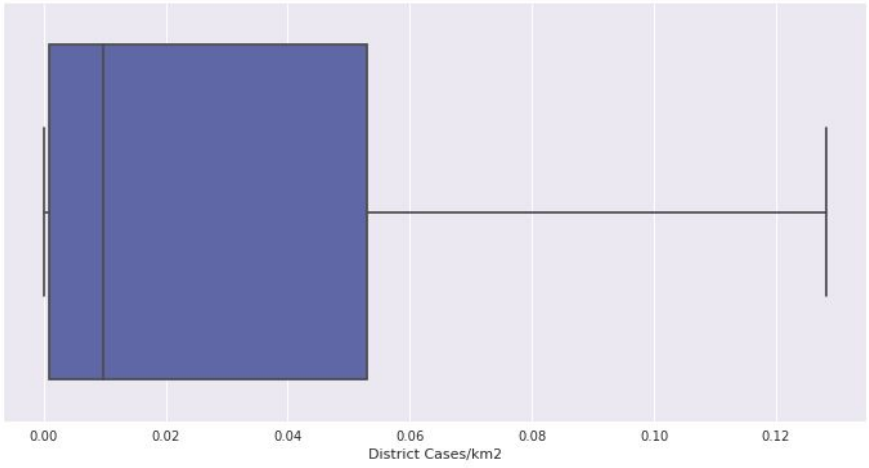
Some exploratory analysis was conducted to decide which features to finally choose for the model.

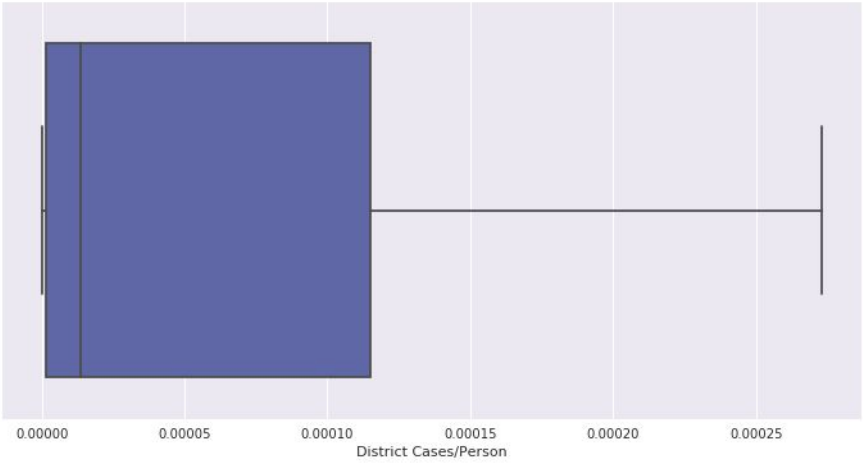






Based on the above three box plots, I decided to take Population Density as a representative rather than the two individual features, since density is more apt for the problem at hand.

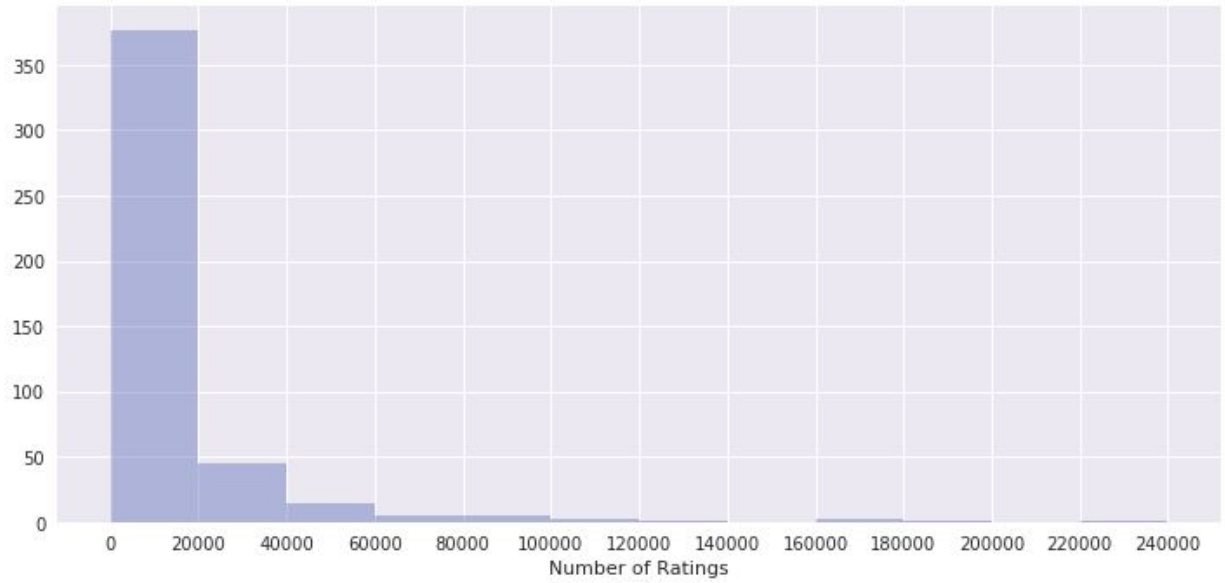




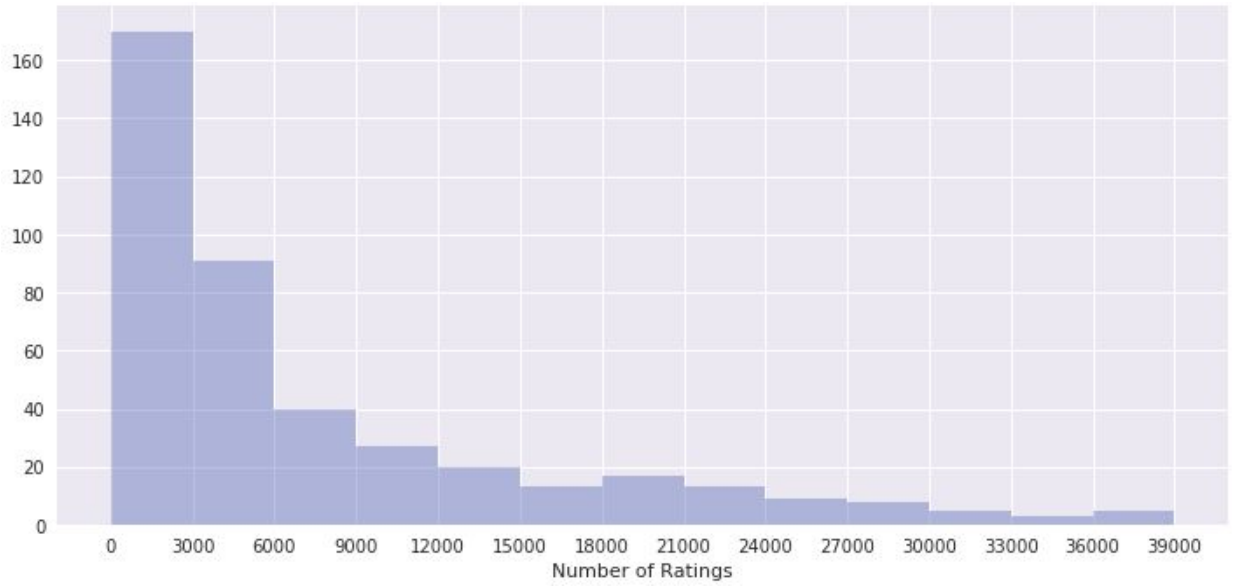
Considering the feature of ‘cases’, I tried both District Cases/Person as well as District Cases/km^2. But both of them were extremely correlated with a correlation coefficient of almost 1 (next image) which made me choose just one for modelling.

	Latitude	Longitude	Ratings	Number of Ratings	District Cases	District Population	District Area	District Cases/Person	District Population Density	District Cases/km2
Latitude	1.000000	0.081925	0.078149	0.095904	-0.029875	-0.069050	0.126111	0.005222	0.072203	-0.018395
Longitude	0.081925	1.000000	0.064979	-0.005431	-0.168116	-0.071701	-0.276664	-0.152946	-0.059064	-0.130376
Ratings	0.078149	0.064979	1.000000	0.234105	-0.005062	-0.164596	0.128691	0.025597	0.011878	0.039115
Number of Ratings	0.095904	-0.005431	0.234105	1.000000	0.128132	0.084905	-0.003490	0.146452	0.153744	0.147309
District Cases	-0.029875	-0.168116	-0.005062	0.128132	1.000000	0.198681	-0.067123	0.968820	0.743097	0.933328
District Population	-0.069050	-0.071701	-0.164596	0.084905	0.198681	1.000000	0.233270	0.050526	0.154823	0.000754
District Area	0.126111	-0.276664	0.128691	-0.003490	-0.067123	0.233270	1.000000	-0.125132	-0.309923	-0.142788
District Cases/Person	0.005222	-0.152946	0.025597	0.146452	0.968820	0.050526	-0.125132	1.000000	0.782538	0.988043
District Population Density	0.072203	-0.059064	0.011878	0.153744	0.743097	0.154823	-0.309923	0.782538	1.000000	0.778164
District Cases/km2	-0.018395	-0.130376	0.039115	0.147309	0.933328	0.000754	-0.142788	0.988043	0.778164	1.000000

A histogram for Number of Ratings was made to understand the variance of the data. From the first image, it can be seen that most of the data fall in the 0-20,000 ratings bracket while the rest is spread over a large range. The second image zooms in on the lower bracket.

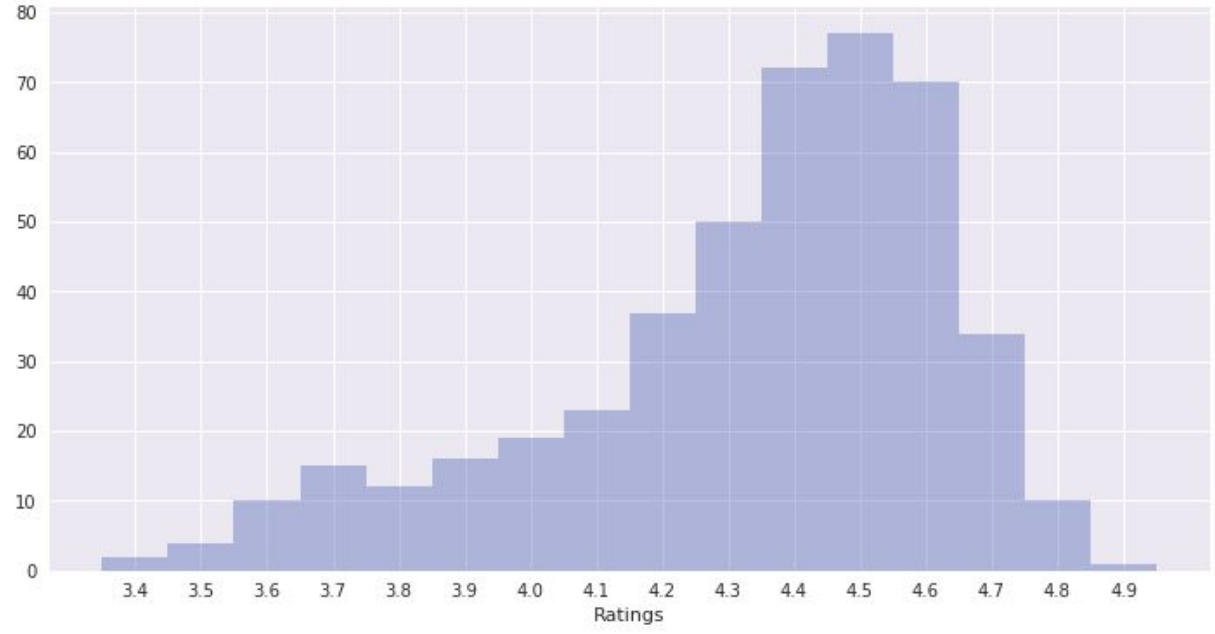


The second image tells us that it is not as bad as the first image depicts, because there are only a few data points (pseudo outliers) that have a very large number of ratings

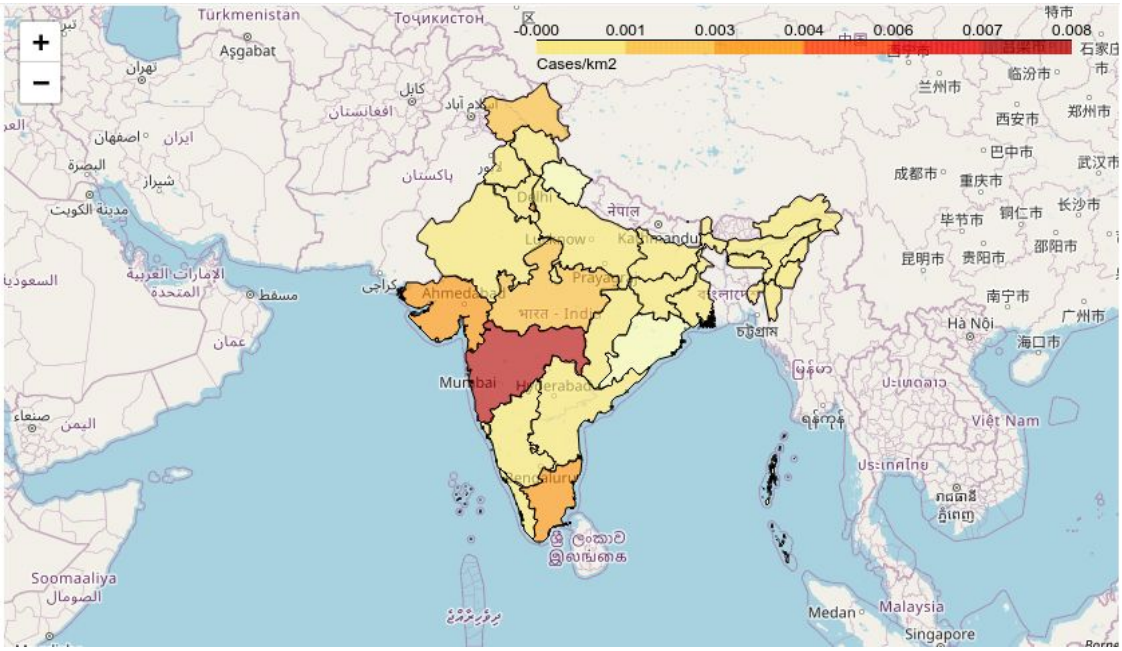


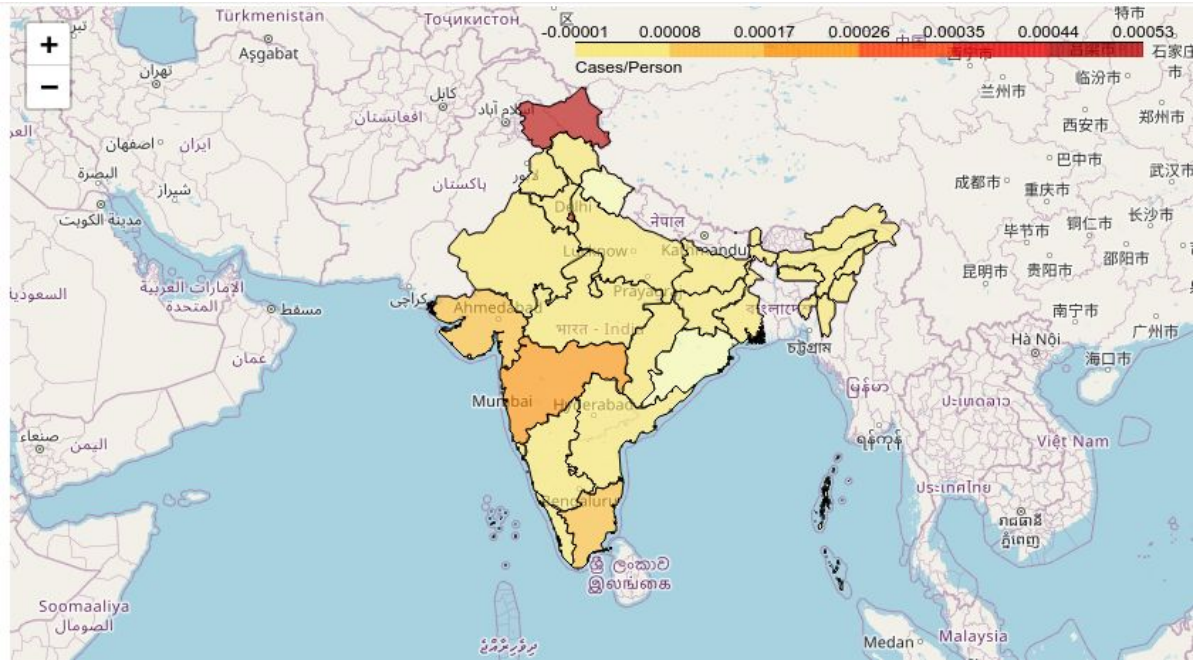
Considering the above as well as the extreme values present in other features like Population density, a standard scaler will be more suitable rather than, say, a min-max scaler.

A histogram of the Rating feature was also plotted.



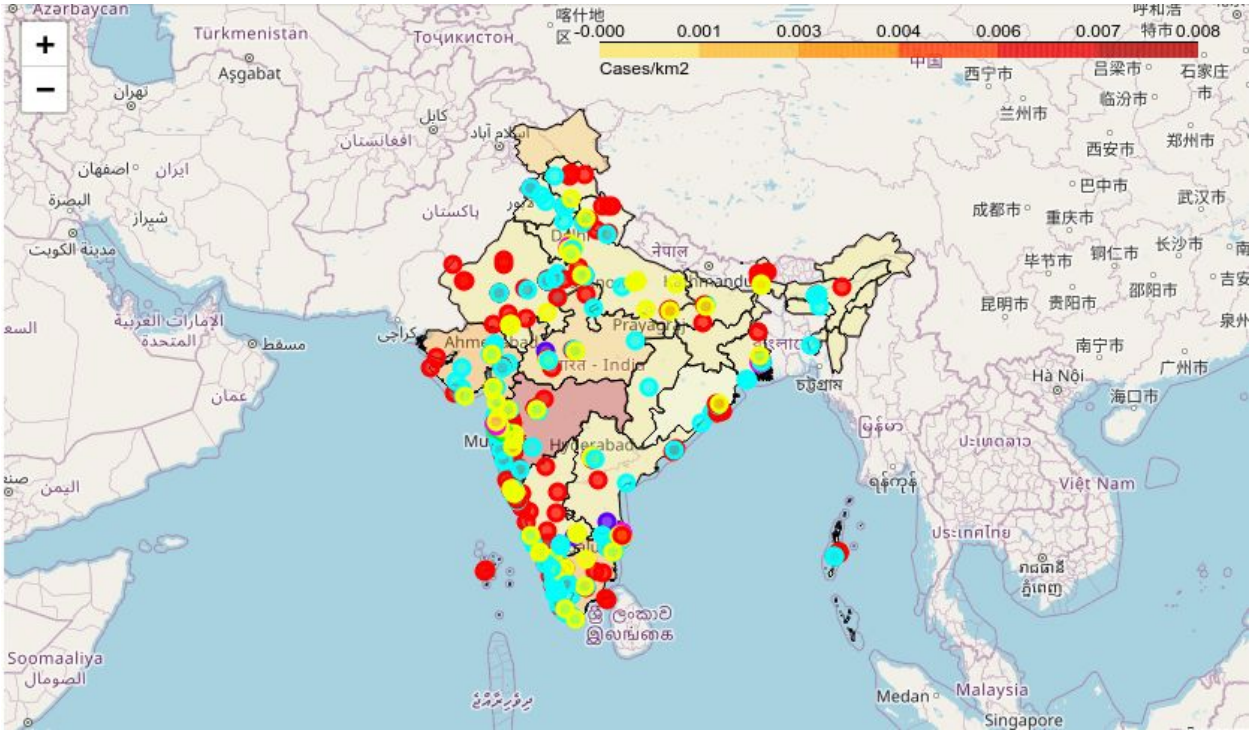
Two choropleth maps for the above two features were also made to get a better visual understanding. One of these, i.e., cases/km^2 was chosen in the results section for the visual plot.



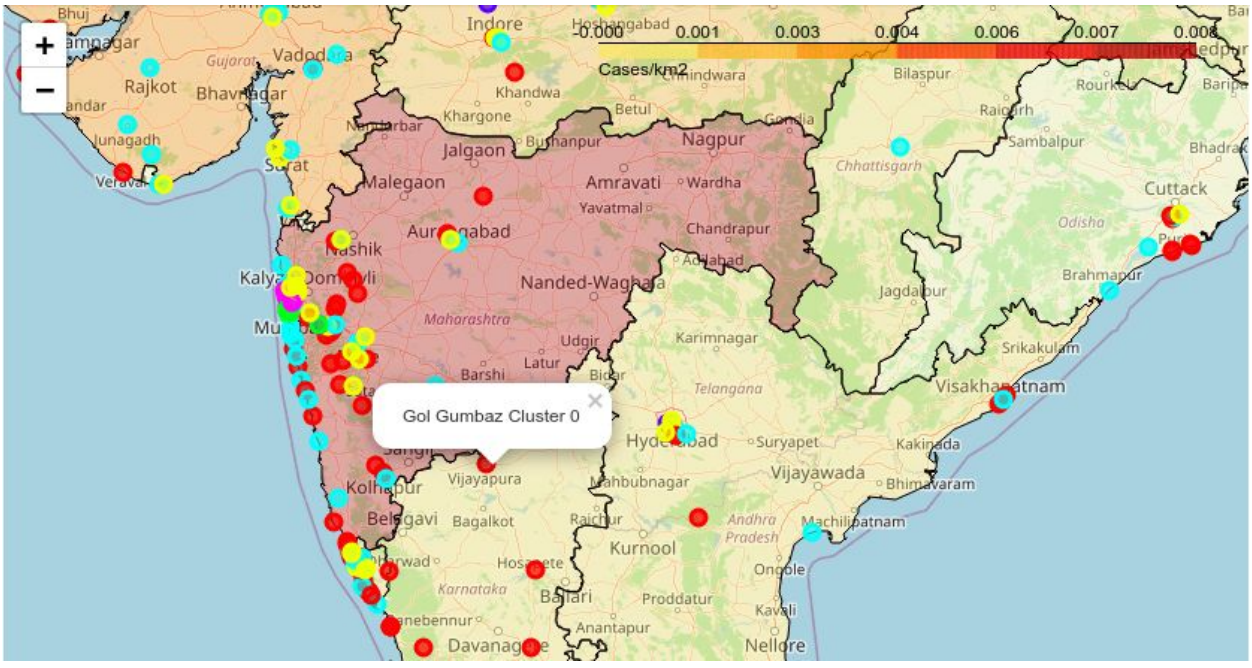


Results

The following plot was obtained after modelling using K-Means clustering.



Zoomed In:



After a few trial and errors, I decided to go ahead with 7 clusters (from 0 to 6), with one cluster, cluster 6, combined with cluster 2, since cluster 6 consisted of a single data point, which was an extreme outlier which related more closely with data points in cluster 2.

Discussion

Using the groupby method to understand the clusters further:

Clusters_KMeans	Ratings	Number of Ratings	District Cases	District Population Density	District Cases/km2
0	4.579769	11914.017341	197.739884	1064.040130	0.109965
1	3.797297	4401.040541	468.662162	1604.429431	0.144802
2	4.388889	14965.111111	14599.000000	45593.710145	211.579710
3	4.288591	5437.201342	272.677852	1118.880619	0.115682
4	4.489474	96278.631579	277.210526	2601.923964	0.192410
5	4.344444	13844.259259	1278.222222	22207.620029	4.482561
6	4.600000	228143.000000	14599.000000	45593.710145	211.579710

The following 6 categories were drawn up to fit all of these clusters:

Cluster 0 (Red) - Very high rating, moderately popular, least crowded and least number of cases - Highly recommended!

Cluster 1 (Yellow) - Very low rating, not popular, moderately crowded and moderate number of cases - Recommended for safety but not hugely for enjoyment.

Cluster 3 (Light blue) - Medium rating, not popular, not very crowded and medium number of cases - Recommended.

Cluster 4 (Dark blue) - High rating, highly popular, moderately crowded and moderate number of cases - Moderately recommended.

Cluster 5 (Pink) - Medium rating, moderately popular, highly crowded and high number of cases - Not recommended.

Cluster 2 (Green) - High rating, moderately popular, extremely crowded and extremely high number of cases - Avoid at all costs.

Possible way forwards:

1. We can further use a district choropleth rather than one which looks at just states. This would help us understand which districts would be more suitable to travel through. This would take a little more time as updated GeoJSON files are not readily available and many districts are called by different names.

2. We can further divide tourist attractions based on type of destination to present it to stakeholders who care only about a certain type.
3. Further zoom in on a state (and perhaps neighbouring states), and conduct a similar deeper analysis for a particular state. This would seem like the next best thing to do considering that each state in India is massive in terms of touristic destinations.
4. Cost can be another feature to be included since it plays a vital role in deciding the right location to travel to.
5. APIs from tripadvisor, which has more wholesome travelling metrics could be used (Unfortunately, they do not allow its use for academia).
6. With coronavirus on the constant rise, it also gives us an opportunity to keep updating this project (can be easily done through the code in the notebook) as and when coronavirus numbers change continuously.
7. With better data, a density based clustering system can be incorporated.

Conclusion

This idea for the project was based on what the future of travelling could look like. With the world in a state of confusion, this project could serve as a data based solution/guide to those wanderlust travelholics who will now look towards domestic tourism. Travelling to many is a way to understand themselves better and learn things you just cannot in a classroom. This project shows that options are available even if safety is our top priority, **it is just a matter of finding those options**.